18 When Philosophers Encounter Artificial Intelligence

The American Academy of Arts and Sciences publishes the journal Dædalus, *with each issue designed, like a camel, by a committee. In 1988, an issue was devoted to Artificial Intelligence, and Hilary Putnam and I both served on the planning committee. I had not intended to contribute an essay to the issue, but when Putnam's essay came in, I decided I had to balance it with one of my own. His view of AI was pure contempt: "What's all the fuss about now? Why a whole issue of* Dædalus*? Why don't we wait until AI achieves something and then have an issue?" There are few philosophers of mind from whom I have learned more than Hilary Putnam. Indeed, his classic papers in the field, beginning with "Minds and Machines" in 1961, were a major inspiration for me, as well as a source of guidance. But in one of his characteristic abrupt shifts of outlook, Putnam turned his back on the ideas of functionalism that he had done so much to clarify. He also turned his back on the nonphilosophical literature that began to accumulate, so that his view of Artificial Intelligence in 1988 did not show much sign of current familiarity with the field. I wanted to make that clear, and also to shed some light on how philosophers might think more constructively about both the strengths and weaknesses of AI.*

How is it possible for a physical thing—a person, an animal, a robot—to extract knowledge of the world from perception and then exploit that knowledge in the guidance of successful action? That is a question with which philosophers have grappled for generations, but it could also be taken to be one of the defining questions of Artificial Intelligence. AI is, in large measure, philosophy. It is often directly concerned with instantly recognizable philosophical questions: What is mind? What is meaning? What is reasoning and rationality? What are the nec-

Originally appeared in *Dædalus*, Proceedings of the American Academy of Arts and Sciences, 117 (1), Winter 1988, pp. 283–295. Reprinted with permission.

VISIT…

LANZAROTE
Caliente.COM

essary conditions for the recognition of objects in perception? How are decisions made and justified?

Some philosophers have appreciated this aspect of AI, and a few have even cheerfully switched fields, pursuing their philosophical quarries through thickets of LISP.[1] In general, however, philosophers have not welcomed this new style of philosophy with much enthusiasm. One might suppose that this is because they have seen through it. Some philosophers have indeed concluded, after cursory inspection of the field, that in spite of the breathtaking pretension of some of its publicists, Artificial Intelligence has nothing new to offer philosophers beyond the spectacle of ancient, well-drubbed errors replayed in a glitzy new medium. And other philosophers are so sure this must be so that they haven't bothered conducting the cursory inspection. They are sure the field is dismissable on "general principles."

Philosophers have been dreaming about AI for centuries. Hobbes and Leibniz, in very different ways, tried to explore the implications of the idea of breaking down the mind into small, ultimately mechanical, operations. Descartes even anticipated the Turing test (Alan Turing's much-discussed proposal of an audition of sorts for computers, in which the computer's task is to convince the judges that they are conversing with a human being [Turing, 1950]) and did not hesitate to issue a confident prediction of its inevitable result:

> It is indeed conceivable that a machine could be made so that it would utter words, and even words appropriate to the presence of physical acts or objects which cause some change in its organs; as, for example, if it was touched in some spot that it would ask what you wanted to say to it; if in another, that it would cry that it was hurt, and so on for similar things. But it could never modify its phrases to reply to the sense of whatever was said in its presence, as even the most stupid men can do. (Descartes, 1637, pp. 41–42)

Descartes's appreciation of the powers of mechanism was colored by his acquaintance with the marvelous clockwork automata of his day. He could see very clearly and distinctly, no doubt, the limitations of that technology. Not even a thousand tiny gears—not even ten thousand—would ever permit an automaton to respond gracefully and rationally! Perhaps Hobbes or Leibniz would have been less confident of this point, but surely none of them would have bothered wondering about the a priori limits on a million tiny gears spinning millions of

[1] The programming language LISP, created by John McCarthy, is the lingua franca of AI.

times a second. That was simply not a thinkable thought for them. It was unthinkable then, not in the familiar philosophical sense of appearing self-contradictory ("repugnant to reason"), or entirely outside their conceptual scheme (like the concept of a neutrino), but in the more workaday but equally limiting sense of being an idea they would have had no way to take seriously. When philosophers set out to scout large conceptual domains, they are as inhibited in the paths they take by their sense of silliness as by their insights into logical necessity. And there is something about AI that many philosophers find off-putting—if not repugnant to reason, then repugnant to their aesthetic sense.

This clash of vision was memorably displayed in a historic debate at Tufts University in March of 1978, staged, appropriately, by the Society for Philosophy and Psychology. Nominally a panel discussion on the foundations and prospects of Artificial Intelligence, it turned into a tag-team rhetorical wrestling match between four heavyweight ideologues: Noam Chomsky and Jerry Fodor attacking AI, and Roger Schank and Terry Winograd defending. Schank was working at the time on programs for natural language comprehension, and the critics focused on his scheme for representing (in a computer) the higgledy-piggledy collection of trivia we all know and somehow rely on when deciphering ordinary speech acts, allusive and truncated as they are. Chomsky and Fodor heaped scorn on this enterprise, but the grounds of their attack gradually shifted in the course of the match. It began as a straightforward, "first principles" condemnation of conceptual error—Schank was on one fool's errand or another—but it ended with a striking concession from Chomsky: it just might turn out, as Schank thought, that the human capacity to comprehend conversation (and more generally, to think) was to be explained in terms of the interaction of hundreds or thousands of jerry-built gizmos—pseudo-representations, one might call them—but that would be a shame, for then psychology would prove in the end not to be "interesting." There were only two interesting possibilities, in Chomsky's mind: psychology could turn out to be "like physics"—its regularities explainable as the consequences of a few deep, elegant, inexorable laws—or psychology could turn out to be utterly lacking in laws—in which case the only way to study or expound psychology would be the novelist's way (and he much preferred Jane Austen to Roger Schank, if that were the enterprise).

A vigorous debate ensued among the panelists and audience, capped by an observation from Chomsky's Massachusetts Institute of Technology colleague, Marvin Minsky, one of the founding fathers of AI, and founder of MIT's AI Lab: "I think only a humanities professor at MIT could be so oblivious to the third interesting possibility: psychology could turn out to be like engineering."

Minsky had put his finger on it. There is something about the prospect of an engineering approach to the mind that is deeply repugnant to a certain sort of humanist, and it has little or nothing to do with a distaste for materialism or science. Witness Chomsky's physics-worship, an attitude he shares with many philosophers. The days of Berkeleyan idealism and Cartesian dualism are over (to judge from the current materialistic consensus among philosophers and scientists), but in their place there is a widespread acceptance of what we might call Chomsky's fork: there are only two appealing ("interesting") alternatives.

On the one hand, there is the dignity and purity of the Crystalline Mind. Recall Aristotle's prejudice against extending earthly physics to the heavens, which ought, he thought, to be bound by a higher and purer order. This was his one pernicious legacy, but now that the heavens have been stormed, we appreciate the beauty of universal physics, and can hope that the mind will be among its chosen "natural kinds," not a mere gerrymandering of bits and pieces.

On the other hand, there is the dignity of ultimate mystery, the Inexplicable Mind. If our minds can't be Fundamental, then let them be Anomalous. A very influential view among philosophers in recent years has been Donald Davidson's "anomolous monism," the view that while the mind *is* the brain, there are *no* lawlike regularities aligning the mental facts with the physical facts (Davidson, 1970). His Berkeley colleague, John Searle, has made a different sort of mystery of the mind: the brain, thanks to some unspecified feature of its biochemistry, has some terribly important—but unspecified—"bottom-up causal powers" that are entirely distinct from the mere "control powers" studied by AI.

One feature shared by these otherwise drastically different forms of mind-body materialism is a resistance to Minsky's tertium quid: in between the Mind as Crystal and the Mind as Chaos lies the Mind as Gadget, an object which one should not expect to be governed by "deep," mathematical laws, but nevertheless a *designed* object, analyzable in functional terms: ends and means, costs and benefits, elegant solutions on the one hand, and on the other, shortcuts, jury-rigs, and cheap ad hoc fixes.

This vision of the mind is resisted by many philosophers despite being a straightforward implication of the current received view among scientists and science-minded humanists of our place in nature: we are biological entities, designed by natural selection, which is a tinker, not an ideal engineer. Computer programmers call an ad hoc fix a "kludge" (it rhymes with Scrooge), and the mixture of disdain and begrudged admiration reserved for kludges parallels the biologists' bemusement with the "panda's thumb" and other fascinating examples of bricolage, to use François Jacob's term (1977). The finest inadvertent spoonerism I ever heard was uttered by the linguist Barbara Partee, in heated criticism of an acknowledged kludge in an AI natural language parser: "That's so *odd hack!*" Nature is full of odd hacks, many of them perversely brilliant. Although this fact is widely appreciated, its implications for the study of the mind are often found repugnant by philsophers, since their traditional aprioristic methods of investigating the mind are relatively powerless to explore phenomena that *may* be contrived of odd hacks. There is really only one way to study such possibilities: with the more empirical mind-set of "reverse engineering."

The resistance is clearly manifested in Hilary Putnam's essay in *Dædalus* (1988), which can serve as a convenient (if not particularly florid) case of the syndrome I wish to discuss. Chomsky's fork, the Mind as Crystal or Chaos, is transformed by Putnam into a pendulum swing he thinks he observes within AI itself. He claims that AI has "wobbled" over the years between looking for the Master Program and accepting the slogan "Artificial Intelligence is one Damned Thing after Another." I have not myself observed any such wobble in the field over the years, but I think I know what he is getting at. Here, then, is a different perspective on the same issue.

Among the many divisions of opinion within AI there is a faction (sometimes called the logicists) whose aspirations suggest to me that they are Putnam's Searchers for the Master Progam. They were more aptly caricatured recently by a researcher in AI as Searchers for "Maxwell's Equations of Thought." Several somewhat incompatible enterprises within the field can be lumped together under this rubric. Roughly, what they have in common is the idea not that there must be a Master Program, but that there must be something more like a master programming language, a single, logically sound system of explicit representation for all the knowledge residing in an agent (natural or artificial). Attached to this library of represented facts (which can be treated as axioms, in effect), and operating upon it computationally, will be one sort or another of "inference engine," capable of deducing the relevant implications of the relevant axioms, and eventually spewing up by this inference process the imperatives or decisions that will forthwith be implemented.

For instance, suppose perception yields the urgent new premise (couched in the master programming language) that the edge of a precipice is fast approaching; this should provoke the inference engine to call up from memory the appropriate stored facts about cliffs, gravity, acceleration, impact, damage, the paramount undesirability of such damage, and the likely effects of putting on the brakes or continuing apace. Forthwith, onę hopes, the engine will deduce a theorem to the effect that halting is called for, and straightaway it will halt.

The hard part is designing a system of this sort that will actually work well in real time, even allowing for millions of operations per second in the inference engine. Everyone recognizes this problem of real-time adroitness; what sets the logicists apart is their conviction that the way to solve it is to find a truly perspicuous vocabulary and logical form for the master language. Modern logic has proven to be a powerful means of exploring and representing the stately universe of mathematics; the not unreasonable hope of the logicists is that the same systems of logic can be harnessed to capture the hectic universe of agents making their way in the protean macroscopic world. If you get the axioms and the inference system just right, they believe, the rest should be easy. The problems they encounter have to do with keeping the number of axioms down for the sake of generality (which is a must), while not requiring the system to waste time re-deducing crucial intermediate-level facts every time it sees a cliff.

This idea of the Axiomatization of Everyday Reality is surely a philosophical idea. Spinoza would have loved it, and many contemporary philosophers working in philosophical logic and the semantics of natural language share at least the goal of devising a rigorous logical system in which every statement, every thought, every hunch and wonder, can be unequivocally expressed. The idea wasn't reinvented by AI; it was a gift from the philosophers who created modern mathematical logic: George Boole, Gottlob Frege, Alfred North Whitehead, Bertrand Russell, Alfred Tarski, and Alonzo Church. Douglas Hofstadter calls this theme in AI the Boolean Dream (Hofstadter, 1985, chap. 26, pp. 631–665). It has always had its adherents and critics, with many variations.

Putnam's rendering of this theme as the search for the Master Program is clear enough, but when he describes the opposite pole, he elides our two remaining prospects: the Mind as Gadget and the Mind as Chaos. As he puts it, "If AI is 'One Damned Thing after Another,' the number of 'damned things' the Tinker may have thought of could be astronomical. The upshot is pessimistic indeed: if there is no Master Program, then we may never get very far in terms of simulating human intelligence." Here Putnam elevates a worst-case possibility (the gadget will be totally, "astronomically" ad hoc) as the only likely alternative to the Master Program. Why does he do this? What does he have against exploring the vast space of engineering possibilities in between Crystal and Chaos? Biological wisdom, far from favoring his pessimism, holds out hope that the mix of elegance and Rube Goldberg found elsewhere in Nature (in the biochemistry of reproduction, for instance) will be discernible in mind as well.

There are, in fact, a variety of very different approaches being pursued in AI by those who hope the mind will turn out to be some sort of gadget or collection of partially integrated gadgets. All of these favor austerity, logic, and order in some aspects of their systems, and yet exploit the peculiar utility of profligacy, inconsistency, and disorder in other aspects. It is not that Putnam's two themes don't exist in AI, but that by describing them as exclusive alternatives, he imposes a procrustean taxonomy on the field that makes it hard to discern the interesting issues that actually drive the field.

Most AI projects are explorations of *ways things might be done,* and as such are more like thought experiments than empirical experiments. They differ from philosophical thought experiments not primarily in their content, but in their methodology: they replace some—not all—of the "intuitive," "plausible" hand-waving background assumptions of philosophical thought experiments by constraints dictated by the demand that the model be made to run on the computer. These constraints of time and space, and the exigencies of specification can be traded off against each other in practically limitless ways, so that new "virtual machines" or "virtual architectures" are imposed on the underlying serial architecture of the digital computer. Some choices of trade-off are better motivated, more realistic or plausible than others, of course, but in every case the constraints imposed serve to discipline the imagination—and hence the claims—of the thought-experimenter. There is very little chance that a philosopher will be surprised (and more pointedly, disappointed) by the results of his own thought experiment, but this happens all the time in AI.

A philosopher looking closely at these projects will find abundant grounds for skepticism. Many seem to be based on forlorn hopes, or misbegotten enthusiasm for one architectural or information-handling feature or another, and if we extrapolate from the brief history of the field, we can be sure that most of the skepticism will be vindicated sooner or later. What makes AI an improvement on earlier philosophers' efforts at model sketching, however, is the manner in which skepticism is vindicated: by the demonstrated, concrete failure of the system in question. Like philosophers, researchers in AI greet each new proposal with intuitive judgments about its prospects, backed up by more or less a priori arguments about why a certain feature *has* to be there, or *can't* be made to work. But unlike philosophers, these researchers are not content with their arguments and intuitions; they leave themselves some room to be surprised by the results, a surprise that could only be provoked by the demonstrated, unexpected power of the actually contrived system in action.

Putnam surveys a panoply of problems facing AI: the problems of induction, of discerning relevant similarity, of learning, of modeling background knowledge. These are all widely recognized problems in AI, and the points he makes about them have all been made before by people in AI, who have then gone on to try to address the problems with various relatively concrete proposals. The devilish difficulties he sees facing traditional accounts of the process of induction, for example, are even more trenchantly catalogued by John Holland, Keith Holyoak, Richard Nisbett, and Paul Thagard in their recent book, *Induction* (1986), but their diagnosis of these ills is the preamble for sketches of AI models designed to overcome them. Models addressed to the problems of discerning similarity and mechanisms for learning can be found in abundance. The SOAR project of John Laird, Allen Newell, and Paul Rosenbloom is an estimable example. And the theme of the importance—and difficulty—of modeling background knowledge has been ubiquitous in recent years, with many suggestions for solutions under investigation. Now perhaps they are all hopeless, as Putnam is inclined to believe, but one simply cannot tell without actually building the models and testing them.

That is not strictly true, of course. When an a priori refutation of an idea is sound, the doubting empirical model builder who persists despite the refutation will sooner or later have to face a chorus of "We told you so!" That is one of the poignant occupational hazards of AI. The rub is how to tell the genuine a priori impossibility proofs from the failures of imagination. The philosophers' traditional answer is: more a priori analysis and argument. The AI researchers' answer is: build it and see.

Putnam offers us a striking instance of this difference in his survey of possibilities for tackling the problem of background knowledge. Like Descartes, he manages to imagine a thought-experiment fiction that is now becoming real, and like Descartes, he is prepared to dismiss it in advance. One could, Putnam says,

> simply try to program into the machine *all* the information a sophisticated human inductive judge has (including implicit information). At the least this would require generations of researchers to formalize this information (probably it could not be done at all, because of the sheer quantity of information involved); and it is not clear that the result would be more than a gigantic expert system. No one would find this very exciting; and such an "intelligence" would in all likelihood be dreadfully unimaginative . . . (1988, p. 277)

This almost perfectly describes Douglas Lenat's enormous CYC project (Lenat, et al., 1986, pp. 65–85). One might say that Lenat is attempting to create the proverbial walking encyclopedia: a mind-ful of common sense knowledge in the form of a single data base containing *all the facts* expressed—or tacitly presupposed—in an encyclopedia! This will involve handcrafting millions of representations in a single language (which must eventually be unified—no small task), from which the inference engine is expected to be able to deduce whatever it needs as it encounters novelty in its world: for instance, the fact that people in general prefer not to have their feet cut off, or the fact that sunbathers are rare on Cape Cod in February.

Most of the opinion setters in AI share Putnam's jaundiced view of this project: it is not clear, as Putnam says, that it will do anything that teaches us anything about the mind; in all likelihood, as he says, it will be dreadfully unimaginative. And many would go further, and insist that its prospects are so forlorn and its cost so great that it should be abandoned in favor of more promising avenues. (The current estimate is measured in person-*centuries* of work, a figure that Putnam may not have bothered imagining in detail.) But the project is funded, and we shall see.

What we see here is a clash of quite fundamental methodological assumptions. Philosophers are inclined to view AI projects with the patronizing disdain one reserves for those persistent fools who keep trying to square the circle or trisect the angle with compass and straightedge: we have *proved* that it cannot be done, so drop it! But the proofs are not geometric; they are ringed with assumptions about "plausible" boundary conditions and replete with idealizations that may prove as irrelevant here as in the notorious aerodynamicists' proofs that bumblebees cannot fly.

But still one may well inquire, echoing Putnam's challenge, whether AI has taught philosophers anything of importance about the mind *yet.* Putnam thinks it has not, and supports his view with a rhetorically curious indictment: AI has utterly failed, over a quarter century, to solve problems that philosophy has utterly failed to solve over two millennia. He is right, I guess, but I am not impressed.[2] It is as if a philosopher were to conclude a dismissal of contemporary biology by saying that the biologists have not so much as asked the question: What is Life? Indeed they have not; they have asked better questions that ought to dissolve or redirect the philosopher's curiosity.

Moreover, philosophers (of all people) should appreciate that solutions to problems are not the only good gift; tough new problems are just as good! Matching Putnam's rhetorical curiosity, I offer as AI's best contribution to philosophy a deep, new, unsolved epistemological problem ignored by generations of philosophers: the frame problem. Plato almost saw it. In the *Theaetetus,* he briefly explored the implications of a wonderful analogy:

Socrates: Now consider whether knowledge is a thing you can possess in that way without having it about you, like a man who has caught some wild birds—pigeons or what not—and keeps them in an aviary he has made for them at home. In a sense, of course, we might say he "has" them all the time inasmuch as he possesses them, mightn't we?

Theatetus: Yes.

Socrates: But in another sense he "has" none of them, though he has got control of them, now that he has made them captive in an enclosure of his own; he can take and have hold of them whenever he likes by

[2] In Dennett (1978a, 1979), I have argued that AI has solved what I called "Hume's Problem": the problem of breaking the threatened infinite regress of homunculi consulting (and understanding) internal representations such as Hume's impressions and ideas. I expect Putnam would claim, with some justice, that it was computer science in general, not AI in particular, that showed philosophy the way to break this regress.

catching any bird he chooses, and let them go again; and it is open to him to do that as often as he pleases. (Cornford trans., 1957, 197C–D)

Plato saw that merely possessing knowledge (like birds in an aviary) is not enough; one must be able to command what one possesses. To perform well, one must be able to get the right bit of knowledge to fly to the edge at the right time (in *real time,* as the engineers say). But he underestimated the difficulty of this trick, and hence underestimated the sort of theory one would have to have to give of the organization of knowledge in order to explain our bird-charming talents. Neither Plato nor any subsequent philosopher, so far as I can see, saw this as in itself a deep problem of epistemology, since the demands of *efficiency* and *robustness* paled into invisibility when compared by the philosophical demand for *certainty,* but so it has emerged in the hands of AI.[3]

Just as important to philosophy as new problems and new solutions, however, is new raw material, and this AI has provided in abundance. It has provided a bounty of *objects to think about*—individual systems in all their particularity that are much more vivid and quirky than the systems I (for one) could dream up in a thought experiment. This is not a trivial harvest. Compare philosophy of mind (the analytic study of the limits, opportunities, and implications of possible theories of the mind) with the literary theory of the novel (the analytic study of the limits, opportunities, and implications of possible novels). One could in principle write excellent literary theory in the absence of novels as exemplars. Aristotle, for instance, could in principle have written a treatise on the anticipated strengths and weaknesses, powers and problems, of different possible types of novels. Today's literary theorist is not *required* to examine the existing exemplars, but they are, to say the least, a useful crutch. They extend the imaginative range, and surefootedness, of even the most brilliant theoretician and provide bracing checks on enthusiastic generalizations and conclusions. The minitheories, sketches, and models of AI may not be great novels, but they are the best we have to date, and just as mediocre novels are often a boon to literary theorists—they wear their deficiencies on their sleeves—so bad theories, failed models, hopelessly confused hunches in AI are a boon to philosophers of mind. But you have to read them to get the benefit.

Perhaps the best example of this currently is the wave of enthusiasm for connectionist models. For years philosophers of mind have been

[3] I present an introduction to the Frame Problem, explaining why it is an epistemological problem, and why philosophers didn't notice it, in chapter 11 of this volume.

vaguely and hopefully waving their hands in the direction of these models—utterly unable to conceive them in detail but sure in their bones that some such thing had to be possible. (My own first book, *Content and Consciousness* [1969] is a good example of such vague theorizing.) Other philosophers have been just as sure that all such approaches were doomed (Jerry Fodor is a good example). Now, at last, we will be able to examine a host of objects in this anticipated class, and find out whose hunches were correct. In principle, no doubt, it could be worked out without the crutches, but in practice, such disagreements between philosophers tend to degenerate into hardened positions defended by increasingly strained arguments, redefinitions of terms, and tendentious morals drawn from other quarters.

Putnam suggests that since AI is first and foremost a subbranch of engineering, it cannot be philosophy. He is especially insistent that we should dismiss its claims of being epistemology. I find this suggestion curious. Surely Hobbes and Leibniz and Descartes were doing philosophy, even epistemology, when they waved their hands and spoke very abstractly about the limits of mechanism. So was Kant, when he claimed to be investigating the conditions under which experience was possible. Philosophers have traditionally tried to figure out the combinatorial powers and inherent limitations of *impressions* and *ideas,* of *petites perceptions, intuitions,* and *schemata.* Researchers in AI have asked similar questions about various sorts of *data structures,* and *procedural representations* and *frames* and *links* and, yes, *schemata,* now rather more rigorously defined. So far as I can see, these are fundamentally the same investigations, but in AI, they are conducted under additional (and generally well-motivated) constraints and with the aid of a host of more specific concepts.

Putnam sees engineering and epistemology as incompatible. I see at most a trade-off: to the extent that a speculative exploration in AI is more abstract, more idealized, less mechanistically constrained, it is "more philosophical"—but that does not mean it is thereby necessarily of more interest or value to a philosopher! On the contrary, it is probably because philosophers have been *too* philosophical—too abstract, idealized, and unconstrained by empirically plausible mechanistic assumptions—that they have failed for so long to make much sense of the mind. AI has not yet solved any of our ancient riddles about the mind, but it has provided us with new ways of disciplining and extending philosophical imagination, which we have only begun to exploit.